Influence of Rule- and Reward-based Strategies on Inferences of Serial Order by Monkeys

Allain-Thibeault Ferhat[1,2] (iD), Greg Jensen[1,2,3],
Herbert S. Terrace[1,2], and Vincent P. Ferrera[1,2]

Abstract

■ Knowledge of transitive relationships between items can contribute to learning the order of a set of stimuli from pairwise comparisons. However, cognitive mechanisms of transitive inferences based on rank order remain unclear, as are relative contributions of reward associations and rule-based inference. To explore these issues, we created a conflict between rule- and reward-based learning during a serial ordering task. Rhesus macaques learned two lists, each containing five stimuli that were trained exclusively with adjacent pairs. Selection of the higher-ranked item resulted in rewards. "Small reward" lists yielded two drops of fluid reward, whereas "large reward" lists yielded five drops. Following training of adjacent pairs, monkeys were tested on novels pairs. One item was selected from each list, such that a ranking rule could conflict with preferences for large rewards. Differences between the corresponding reward magnitudes had a strong influence on accuracy, but we also observed a symbolic distance effect. That provided evidence of a rule-based influence on decisions. RT comparisons suggested a conflict between rule- and reward-based processes. We conclude that performance reflects the contributions of two strategies and that a model-based strategy is employed in the face of a strong countervailing reward incentive. ■

INTRODUCTION

When making choices, individuals often rely on rankings of available options. Ranking may be based on prior experience or on the application of general rules. To predict the outcome of a baseball game between two teams that have never faced one another, one might consider how they fared against common opponents. Given that Team A reliably beats Team B, and Team B beats Team C, it is reasonable to conclude that Team A would likely beat Team C. This is an example of a *transitive inference* (TI), the ability to infer ordinal relationships based only on implied rank without reliance on other cues (Halford, Wilson, & Phillips, 2010; Piaget, 1921). Behavior consistent with TI has been observed in a wide variety of vertebrate species, including rhesus macaques (*Macaca mulatta*; Lazareva, Smirnova, Zorina, & Rayevsky, 2001; Treichler & Van Tilburg, 1996; Davis, 1992; Gillan, 1981; McGonigle & Chalmers, 1977).

Transitivity is a logical feature of ordered sets and can therefore be implemented as a rule for making choices consistent with limited information. However, such rules may conflict with other heuristics, such as preferring items based on a subjective sense of their value. For example, some studies suggest that gamblers tend to bet on their home team even when the odds of winning are substantially less than even (Staněk, 2017). In this study, we investigated how monkeys made decisions when a rule based on rank conflicted with experienced reward values.

In behavioral paradigms that test TI, the experimenter assigns ranks to stimuli to create an ordered list (e.g., A > B > C > D > E). During training, subjects are presented with adjacent pairs from that list (AB, BC, CD, etc.), receiving a reward if they choose the item of superior rank in each pairing (Jensen, 2017). One clue that subjects are performing TI is a preference consistent with the rank order for novel critical test pairs in which both stimuli had equal reward rates during training (e.g., BD). Another is the presence of symbolic distance effects (SDEs; D'Amato & Colombo, 1990), in which performance improves as the gap between items' ranks grows.

A variation of the TI paradigm is to train multiple lists in parallel and then to present new derived lists that are assembled from a mixture of training lists (Hakes, James, & Young, 1964). Monkeys tend to preserve the relative ranks of items presented in new derived combinations (Jensen, Ferrera, & Terrace, 2020; Treichler & Raghanti, 2010; Chen, Swartz, & Terrace, 1997; Swartz, Chen, & Terrace, 1991). This suggests that they rely on linear representations of rank to evaluate the otherwise ambiguous derived pairs (Jensen et al., 2020; Mione, Brunamonti, Pani, Genovesio, & Ferraina, 2020; Merritt & Terrace, 2011).

Although the mechanisms that result in TI remain unclear, reinforcement learning (RL) provides a framework for characterizing potential strategies. Algorithms that make choices only by associating stimuli with expected reward values are *model-free* because they do not rely on any other organizational structure (Jensen, Terrace, & Ferrera, 2019; Dayan & Niv, 2008). Model-free learning therefore depends on reward history and cannot exploit rules based on the assigned ranks of stimuli (Ferrucci, Nougaret, Brunamonti, & Genovesio, 2019; Stanisor, Van Der Togt, Pennartz, & Roelfsema, 2013; Minamimoto, La Camera, & Richmond, 2009). Model-free strategies, including the value transfer model (von Fersen, Wynne, Delius, & Staddon, 1991), fail to completely describe behavior in many TI paradigms (Jensen, Alkan, Ferrera, & Terrace, 2019; Vasconcelos & Monteiro, 2014; Gazes, Chee, & Hampton, 2012; Lazareva & Wasserman, 2006, 2012).

As an alternative approach, *model-based* algorithms assume that stimuli relate to one another within a representational framework and that a subject's choice is based on rules relating to that representation (Dayan & Niv, 2008). Model-based accounts of TI that rely on the representation of an item's position along a continuum predict most TI behaviors, including SDEs (Kao, Jensen, Michaelcheck, Ferrera, & Terrace, 2020; Jensen et al., 2020; Mione et al., 2020). This continuum has previously been described as a cognitive representation of a serial order (Terrace, 2010, 2012).

Jensen, Alkan, et al. (2019) showed that *Q*-learning, a model-free algorithm, cannot solve classic TI, either when all rewards are equal or when low-ranked items are associated with larger rewards than high-ranked items. However, *Q*-learning succeeds when differential rewards are concordant with the ranking rule (Jensen, Alkan, et al., 2019; Jensen, Terrace, et al., 2019). This suggests that model-free RL can be a useful heuristic in some TI contexts, although behavior in TI tasks is otherwise well described by model-based RL (Smith & Church, 2018).

In this study, we measured the contributions of reward-based and rule-based components of learning and decision-making by training subjects using lists for which the magnitude of reward differed. The conflict between reward disparity and relative item rank in derived pairs allowed the dissociation of rule-based and reward-based contributions to preference.

METHODS

Subjects

Subjects were three adult male rhesus macaques (*M. mulatta*) identified as O, R, and S and housed at Columbia University. The study was carried out in accordance with the guidelines provided by Guide for the Care and Use of Laboratory Animal of the National Institutes of Health. The Institutional Animal Care and Use Committee at Columbia University approved the study.

All subjects had previous experience with serial learning procedures, including TI tasks. Subjects earned rewards that were delivered in units of "drops," with each drop having a volume of approximately 0.20 mL. Typical performance during a session earned between 100 and 250 mL, whereas perfect performance could earn up to 300 mL. Subjects also received a ration of biscuits (provided before the session) and fruit (provided after the session).

Apparatus

Subjects performed the task using a touchscreen that was linked to a computer. The subjects sat in a specially designed primate chair while performing the task. The touchscreen (model 1939 L, Elo Touch Solutions) had a 19-in. display (1280 × 1024 resolution at 60 Hz) with a resistive touchscreen interface to record responses. All tasks were programmed in MATLAB (2019, The MathWorks) using the Psychophysics Toolbox (Kleiner et al., 2007; Brainard, 1997; Pelli, 1997). To deliver fluid rewards, the computer was connected to a solenoid valve through an Arduino Uno interface. Fluid was delivered through a stainless-steel tube attached to the primate chair.

Stimuli and Task

Stimuli were presented on the touchscreen. Subjects responded by touching them when they appeared. The stimuli were photographic images drawn from a large set of royalty-free stock images. They were grouped into randomly drawn lists of five images, with an ordering predefined by the experimenter. We denote the item positions using the labels ABCDE, with A being the highest rank and E the lowest rank. No indication of stimulus rank was presented to subjects beyond that implied by correct/incorrect feedback. In addition, stimulus lists were inspected before the experiment to ensure that stimulus features were not predictive of list order. The images were presented in pairs, one image on the left and one on the right side of the display. Only two stimuli were presented on each trial. Stimulus arrangement on screen was counterbalanced over trials to avoid left/right bias.

Figure 1A presents the experimental design. Each trial began with the presentation of a solid blue square (100 × 100 pixels) at the center of the screen to focus the attention of the subject and to provide a consistent target for their hand. After the subject touched the square, two stimuli (200 × 200 pixels) were presented on either side of the screen. These stimuli always differed in rank. Touching the stimulus with the higher rank was "correct" and resulted in a fluid reward and the presentation of a green check as visual feedback. Touching the stimulus with the lower rank was "incorrect" and resulted in the presentation of a red X, which was followed by a time-out lasting 2 sec.
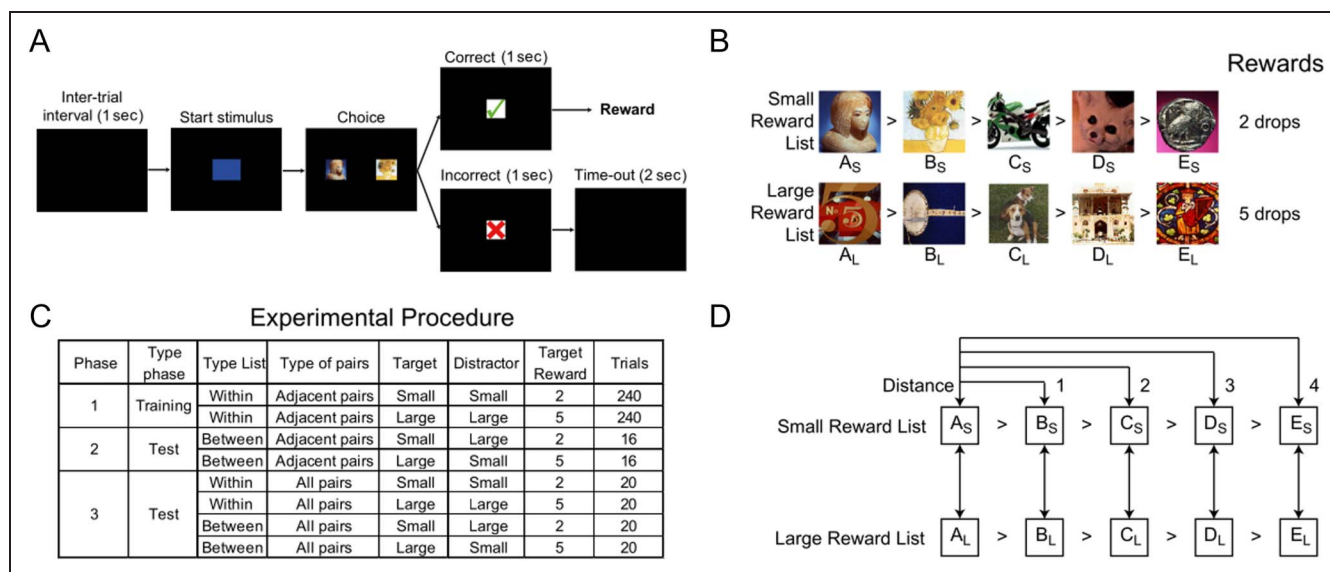
Figure 1. Description of the experimental procedure. (A) Representation of trial procedure. The trial started with the presentation of a blue square after a 1-sec intertrial followed by the presentation of pairs from within or between lists. If the subject touched the higher-ranked item, a reward was delivered according to the type of list associated. Otherwise, a time out of 2 sec occurred. (B) Example of the five-item list annotated small list and large reward list and the reward associated, consisted of randomly picked stock image. (C) A detailed description of the type of pairs presented during each of the three phases. All types of pairs appearing during the same phase were intermixed with one another. (D) Schematic description of the interaction between all items according to the distance and the possibility of switch items between the small reward list and large reward list.

During the time-out, the screen was blank. For example, when presented with the pair AB, a reward could only be earned by touching the stimulus A.

Two 5-item lists (with ranks denoted by ABCDE, e.g., Figure 1B) were used in each experimental session. The 10 images used in these two lists had not been seen in any prior session and were not reused in subsequent sessions. To study the effect of reward magnitude on learning, correct responses on one list were rewarded with five drops of fluid (the "large reward" list); on the other, with two drops (the "small reward" list). We denote which list a stimulus belongs to with a subscript, such that A_S indicates the highest-ranked item in the small reward list, whereas E_L indicates the lowest-ranked item in the large reward list. As a control, there was also an "equal reward" condition, during which subjects still learned two lists but all correct responses yielded rewards of two drops of fluid.

Each subject completed 25 sessions per condition, one session per day. Each session comprised a training phase, followed by two different testing phases. During training, subjects were presented trials with adjacent pairs where the two items were from the same list ("within list" pairs). This yielded eight pairs: A_SB_S, B_SC_S, C_SD_S, D_SE_S, A_LB_L, B_LC_L, C_LD_L, and D_LE_L. Whereas each trial presented only items from one list, the two lists were randomly interleaved across trials. Trials were organized into blocks, during which each pair was presented twice: once with the target (higher ranked item) on the left, and once on the right. The sequence of presentations within a block was randomized. Training continued in this fashion for 30 blocks or 480 trials.

During the first testing phase, a new set of adjacent pairs was assembled by taking one item from each list ("between lists" pairs). This yielded eight new pairs: A_SB_L, B_SC_L, C_SD_L, D_SE_L, A_LB_S, B_LC_S, C_LD_S, and D_LE_S. When evaluating choices made for these derived pairs, the amount of reward was determined by which list the correct item was drawn from. If the subject was presented with the pair A_SB_L, choosing A_S yielded a reward of two drops and choosing B_L yielded a time-out. However, if the subject was presented with the pair A_LB_S, choosing A_L yielded a reward of five drops and choosing B_S yielded a time-out. These eight pairs were also arranged into blocks of trials, such that presentations within each block randomly permuted the order of presentation. Each pair appeared twice per block, once with the target on the left and once with the target on the right. This first testing phase consisted of two blocks, for a total of 32 trials.

During the second testing phase, all possible pairs were presented, including both within-list pairings such as B_SD_S and between-lists pairings such as C_SE_L. Because 10 pairs can be drawn from five items, there were 10 within-list small reward pairs, 10 within-list large reward pairs, 10 between-lists pairs with a small reward target, and 10 between-lists pairs with a large reward target. The presentation order of each pair was randomized, with all pair types intermixed and with counterbalanced target positions. Thus, the second testing phase comprised a total of 80 trials. Figure 1C summarizes these trial counts and compositions. Figure 1D gives examples of each of the four types of pairings that can occur in this experiment overall: adjacent within-list pairs (symbolic distance of 1, e.g., A_SB_S, seen

during training), adjacent between-lists pairs (symbolic position of 1, e.g., $A_L B_S$, seen during the first testing phase), and all-pairs combinations that included both within-list pairs and between-lists pairs (e.g., $A_S E_S$ and $C_S E_L$, seen during the second testing phase).

If the subject failed to initiate a trial by pressing the start stimulus or initiated and then waited more than 4 sec to make a choice, the trial was not counted toward the requirement for completing that phase and was presented again. For the unequal reward condition, within-list adjacent-pairs presentation, an average of 562.98 trials were attempted, relative to the 480 completed trials necessary to complete the phase trials (O: 541.36, R: 664.56, S: 483.04); for between-lists adjacent-pairs presentations, 46.21 trials were attempted of the 32 required (O: 59.64, R: 42.12, S: 36.88); for all-pairs presentations, 100.53 trials were attempted of the 80 required (O: 106.84, R: 105.12, S: 89.64). For the equal reward condition, within-list adjacent-pairs presentations, 562.98.24 trials were attempted of the 480 required (O: 542.72, R: 761.68, S: 481.92); for between-lists adjacent-pairs presentations, 46.21 trials were attempted of the 32 required (O: 42.72, R: 40.24, S: 37.04); for all-pairs presentations, 100.53 trials were attempted of the 80 required (O: 97.48, R: 112.04, S: 89.16).

Analysis

The number of subjects, as well as the number of session, was based on previous work on nonhuman primate, especially on serial learning (Jensen et al., 2020; Jensen, Alkan, et al., 2019; Jensen, Terrace, et al., 2019). Each subject's performance, quantified as proportion of correct responses, was modeled using logistic regression with trial number, symbolic distance, and reward size as explanatory variables. These variables are related to outcomes using the following logistic transformation:

$$p(\text{correct}) = \frac{1}{1 + \exp(-\mu)} \qquad (1)$$

Here, μ represents a linear combination of terms, which varies with the phase of the task being modeled. Unless otherwise specified, individual predictors that contribute to μ use mean-centered values in order to avoid creating strong covariance with intercept terms. Consequently, in all models described below, the intercept (denoted by $\beta_\varnothing$) corresponds to a "typical" case, and each predictor corresponds to some deviation from that typical case.

Because our chief interest was to assess the contributions of both rule-based and reward-based influences on behavior, we relied on a regression model to tease apart these contributions. One of the important rule-based effects we expected to observe was the SDE (D'Amato & Colombo, 1990). When a subject displays an SDE, their response accuracy, or RTs, or both change as a function of the relative rank of the items. For example, given a list

ABCDE, a positive effect of distance would result in BE (a distance of 3) having higher accuracy than BD (a distance of 2). SDEs are frequently observed in the empirical literature and are consistent with the view that TI relies on a cognitive representation of list order (Jensen, 2017; Terrace, 2012). We operationalized D as our measure of symbolic distance in regressions that include pairs of multiple distances. Although the symbolic distance between items varied from 1 to 4, smaller symbolic distances were more numerous than large ones. Because of this, D was centered by subtracting 2 from every case, the mean distance of all possible pairs. For example, the value of D for the pair AB was $-1 = (1 - 2)$, whereas the value of D for the pair AD was $1 = (3 - 2)$. This centering ensured that estimates of β_D did not covary with other parameters any more than necessary.

Similarly, we operationalized our reward manipulation as R, a predictor that was a centered and dummy-coded, with $R = .5$ when the target stimulus belonged to the large reward list and $R = -.5$ when the target belonged to the small reward list.

During training, every trial used adjacent pairs, which have a symbolic distance of 1. Because there was no variation in distance among these pairs, training performance was modeled based on the learning rate in terms of trials t and the reward amount R, as well as the interaction of these two variables. We categorized training data as "early" (first third of trials) and "late" (last third of trials), to allow the model to best fit to data. This yielded the following linear equation during training:

$$\mu = \beta_\varnothing + R \cdot \beta_R + t \cdot \beta_t + R \cdot t \cdot \beta_{Rt} \qquad (2)$$

Here, $\beta_\varnothing$ is an intercept term that describes overall response accuracy at the start of training. As noted above, β_R describes any differential response accuracy observed at the start of training. β_t acts as a learning rate, describing how response accuracy changes over time, and β_{Rt} describes the differential learning rate between the small reward list and the large reward list.

The first testing phase (between-lists adjacent-pairs) consisted of too few trials for a learning rate β_t to be estimated. With this in mind, we used the following linear combination:

$$\mu = \beta_\varnothing + R \cdot \beta_R \qquad (3)$$

Because trials are omitted as a predictor, $\beta_\varnothing$ and β_R are more appropriately interpreted as descriptions of overall performance during the adjacent-pair between-lists testing phase. As above, R was coded as the reward magnitude associated with the target, with a value of .5 when the target belonged to the large reward list and $-.5$ when the target belonged to the small reward list.

During the final testing phase, trials were again omitted as an explanatory variable, and performance was predicted in terms of R, the reward size associated with the target. In addition, the symbolic distance D was also used as a

predictor, resulting in the following linear combination:

$$\mu = \beta_\varnothing + R \cdot \beta_R + D \cdot \beta_D + D \cdot R \cdot \beta_{DR} \qquad (4)$$

As noted above, D was a centered predictor, and as such, estimates of β_D covaried minimally with estimates of β_R or of $\beta_\varnothing$. The predictor β_{DR} describes the interaction between D and R on the overall response accuracy.

During the equal reward condition, the parameters of reward β_R, β_{DR}, and β_{Rt} were removed from the equations described above because R did not vary in that condition. In addition, we treated performance for List 1 and List 2 as interchangeable with respect to targets for equal reward condition data. As such, A_1B_1 and A_2B_2 were both simply treated as within-list pairs, whereas A_1B_2 and A_2B_1 were both treated as between-lists pairs.

Parameters were estimated using multilevel models, with each regression parameter estimated for each subject, as well as at the population level. These models were implemented in the Stan programming language (Stan Development Team, 2021; Carpenter et al., 2017) and scripted using the R language (R Foundation for Statistical Computing).

RESULTS

To study the impact of reward disparity on learning and decision-making, we first trained subjects on two ordered lists of five items each, where each trial presented a pair of items drawn from one list or the other (Figure 1). Each daily session comprised an initial training phase followed immediately by the testing phases. During training, only within-list adjacent pairs were presented, with either small or large reward size associated with each pair for unequal reward condition (Figure 1C). The different possible pairs were then presented during the following test phase (Figure 1C). Each daily session used 10 new items on which subjects had not been trained previously.

During training, performance reached 75% correct responses for both small and large reward lists (Figure 2A). Only the CD pairs showed lower performance compare to other adjacent pairs, a pattern consistent with past reports in the literature (Jensen, Alkan, et al., 2019). Thus, when subjects learned two lists in the same session, reward magnitude did not seem to have an effect on overall performance at the end of training, as long as both items presented on each trial belonged to the same list.

We then observed the performance for three types of test pairs: "adjacent between-lists pairs," "all within-list pairs," and "all between-lists pairs." For test pairs comprising adjacent items drawn from different lists with the same reward magnitude, subjects were able to perform well above chance (Figure 2B, purple line), consistent with previous studies (Terrace, 2010). This establishes that subjects were able to make correct between-lists comparisons when reward was not a confound. We have previously referred to this ability as "positional inference" (Jensen et al., 2020).

However, when a target item from the small reward list was presented with a distractor item from the large reward list, adjacent pair performance was far below chance levels (Figure 2B, orange points). In other words, nonhuman primates tended to choose the item that had been associated with the larger reward during training, although this went against the rule of choosing the item with higher rank. Such choices were not rewarded during testing, but rather resulted in a time-out.

Conversely, in adjacent pairs where the target was from the large reward list and the distractor from the small reward list, performance was not only above chance but also better than performance in the equal reward condition (Figure 2B, green points). All conditions displayed a similar pattern of higher performance for terminal pairs (AB and DE) and lower performance for CD pairs (Figure 2B), despite their different baseline accuracies.

Performance for all combinations of items from both lists (all collected simultaneously during the final testing phase) is here partitioned into within-list pairs (Figure 2C) and between-lists pairs (Figure 2D). This testing phase included pairs with symbolic distances of 1 (adjacent pairs) to 4. In all cases, an SDE was evident, with performance generally increasing as a function of symbolic distance.

Furthermore, for the equal reward condition, within- and between-lists performance was nearly identical, providing further evidence of positional inference. Performance for within-list pairs was also similar when comparing the equal reward condition to the unequal pairs with large reward targets (Figure 2C). Performance for pairs with small reward targets, however, was consistently lowest of the three (Figure 2C, orange line).

As in Figure 2B, performance for between-lists pairs with small reward targets during all-pairs testing was below chance in many cases (Figure 2D). Despite this, subjects still displayed SDEs, and as such, small reward target pairs were selected above chance when the symbolic distance became large enough. Pairs with large reward targets showed the opposite reward effect, with elevated performance relative to the equal reward condition, effectively reaching a performance ceiling.

Figures 2E–2H plot the performance for "critical test pairs," those novel adjacent or nonadjacent pairs that omit both the first and the last items (e.g., BD or CD, but not AC or BE). Performance for pairs with large reward targets was consistently at or above 75% correct and either equal or superior to performance in the equal reward condition. However, small reward target pairs were below chance in all between-lists cases (Figure 2E, 2F, and 2H) and showed a disadvantage even in within-list comparisons, despite mostly being above chance (Figure 2G).

Overall, performance strongly suggest that subjects understood the rule to "choose the higher rank" and were able to do so both when rewards were equal and when they were concordant with corresponding reward
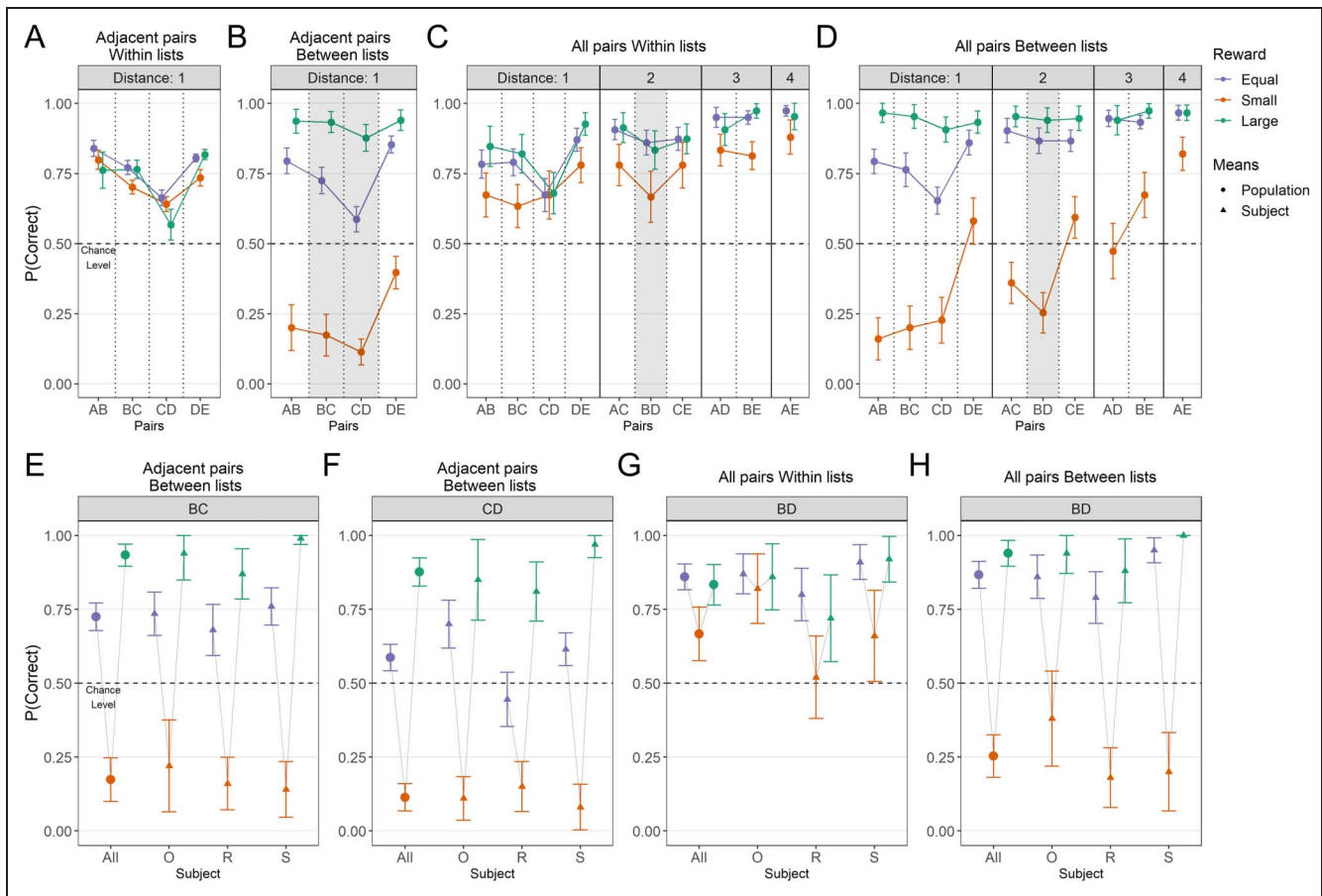
Figure 2. Observed average response accuracy of each type of pairs (A–D) and critical pairs (E–H) for each reward manipulation. Points and error bars depict the average and 95% confident interval of observed performance for each pair. Performance (A) during training, adjacent pairs within list presentation, (B) during adjacent pairs between list presentation, (C) during all pairs within list presentation, (D) and during all-pairs between-lists presentation. (E–H) Performance of critical pairs during equal, the small reward condition, and the large reward condition for the population and each subject: BC and CD pairs during between-lists adjacent-pairs presentations (E and F) and BD during within-list (G) and between-lists all-pairs presentations (H). Pairs are annotated accordingly to type of reward, equal condition (purple), small rewarded target (orange), and large rewarded target (green).

magnitude (Jensen, Alkan, et al., 2019). Furthermore, even when a rule–reward conflict drove performance below chance, the presence of SDEs during all-pair testing suggests that subjects still demonstrated some understanding of the list order, despite the biasing effect of the reward disparity. This selective deployment of a rule-based strategy raises the question of how the rule interacted with reward associations, particularly when the two were in conflict. To address this, we examined RTs.

Figure 3 shows RTs on a log scale for all pairs (Figure 3A–3D) and critical pairs (Figure 3E–3H), split for correct/incorrect responses. In general, during all phases and regardless of whether correct or incorrect, mean RTs for the equal reward condition had the same magnitude (Figure 3A–3D, purple points). An overall SDE is also visible, with faster RTs for large symbolic distances.

However, the introduction of a reward disparity induces a substantial split in RTs. between-lists pairs with small reward targets are, when selected correctly, also chosen substantially more slowly than incorrect responses (Figure 3B and 3D, filled points). Even the within-pair correct trials

of the testing phase show a slowdown for small reward targets when they are paired with small reward distractors (Figure 3C, filled points). In contrast, when the target is associated with the larger reward, RTs for correct responses were comparable to equal reward trials for within-list pairs and faster than equal reward trials among correct between-lists pairs (Figure 3B and 3D, filled points). In summary, when a small reward target was paired with a large reward distractor, correct responses were particularly slow and deliberate, whereas when a large reward target was paired with a small reward distractor, correct responses were particularly fast.

Figure 3E–3H directly contrast the correct and incorrect response RTs for the critical pairs, BC, CD, and BD, doing so for both overall means and for individual subjects. In the equal condition, differences in RTs between correct and incorrect responses were small or nonexistent. For all between-lists means, small reward trials (orange) were slow when the response was correct, and were fast when the response was incorrect (Figure 3E, 3F, and 3H). The opposite was true for the large reward trials (green). This difference in RT suggests an increased difficulty and a

cognitive cost to respond during decisions in which the rule and the reward were in conflict. A competition between rule- and reward-based decision processes, rather than just an overall tendency to respond more quickly to items associated with larger rewards, seems to explain this result. Note that because subject S had a 100% correct response rate to large reward targets in between-lists pairs (Figure 2H), we cannot estimate that subject's RT for incorrect responses (Figure 3H).

To quantify the relative contributions of rule- and reward-based decision processes, we fitted performance data with the regression model described in Equations 1–4 to each phase. This regression revealed three aspects of the learning: (a) $\beta_\varnothing$, the performance at the beginning of each phase (or the phase overall when no learning rate was included); (b) β_D, the impact of increased distance between pairs on performance, associated with the SDE and the rule-based decision process; (c) β_R, the impact of reward disparity on performance, a proxy of the contribution of reward-based process; and (d) the relation between β_D and β_R during all pairs presentation. Figure 4A and 4C show model parameter estimates for the equal reward condition in each phase of the experiment, whereas Figure 4B, 4D, 4E, and 4F do the same for the
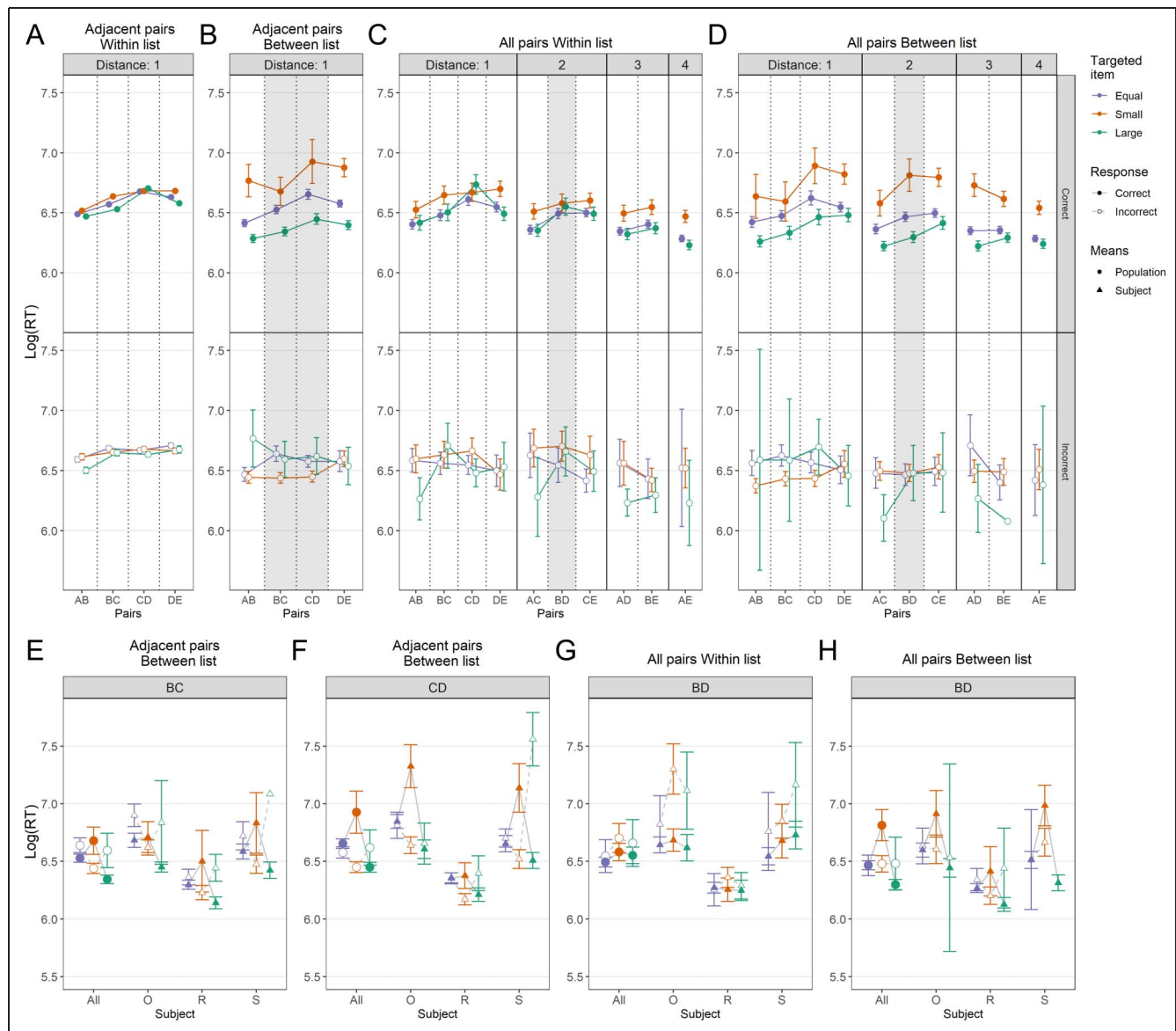

Figure 3. Observed average RT of each type of pair (A–D) and critical pairs (E–H) for each subject and each reward manipulation. Filled points correspond to RTs for correct responses, whereas open points correspond to RTs for incorrect responses. Points and error bars depict the average and 95% confident interval of observed RT for each pair. Mean log RTs (A) during training, adjacent pairs within list presentation, (B) during adjacent pairs between list presentation, (C) during all pairs within list presentation, (D) during all-pairs between-lists presentation. (E and F) Mean log RTs for the pairs BC and CD during between-lists adjacent-pairs presentations for equal, the small reward target, and the large reward target. (G and H) RTs for the pairs BD during within-list and between-lists all-pairs presentations. Pairs are annotated accordingly to type of reward, equal condition (purple), small rewarded target (orange), and large rewarded target (green).
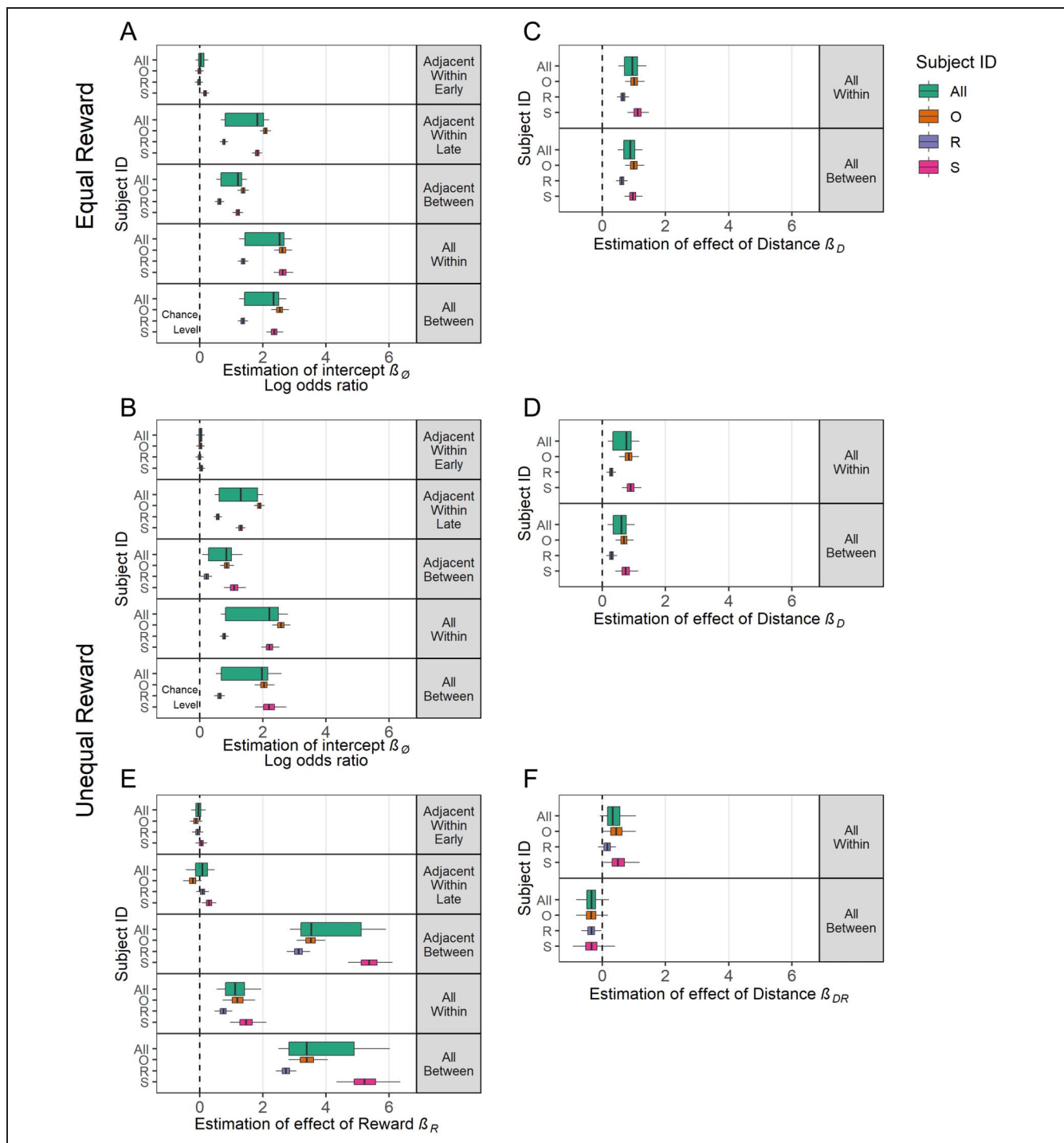
Figure 4. Estimated logistic regression parameters for the population (green box) and each subject (O = orange, R = purple, and S = pink) for equal and unequal reward manipulation. Boxes represent the 80% credible interval, and whiskers represent the 95% credible interval. (A and B) Estimates of the intercept $\beta_{\varnothing}$, describing performance at trial 0 for each type of pair presentation for the equal reward condition (A) and unequal reward condition (B). (C and D) Estimates of β_D, describing the impact of symbolic distance for the equal reward (C) and unequal reward (D). (E) Estimates of β_R, describing the impact of reward disparity for the unequal reward condition. (F) Estimates of β_{DR}, describing the interaction between D and R on the overall response accuracy for the unequal reward condition.

unequal reward condition. At the beginning of each session ("Adjacent Within Early"), the intercept $\beta_{\varnothing}$ was at chance levels for each subject in each condition (unequal and equal reward, Figure 4A and 4B "Adjacent Within Early"). During late-stage training and all subsequent test phases, values of $\beta_{\varnothing}$ were consistently positive, demonstrating that subjects performed above chance overall, whether there was a reward disparity or not. Performance

that is reliably above chance on trial 0 indicates that information has been learned and retained from the previous phase of the experiment.

The estimates of $\beta_\varnothing$ for "Adjacent Between" reflect performance during between-lists testing. Values above zero indicate that the representation of order learned during training transferred to testing, insofar as positional inferences among adjacent pairs are concerned. $\beta_\varnothing$ was lower for the unequal rewards condition than for the equal rewards condition, suggesting that the reward disparity in some way tended to impair transfer.

The predictor β_D captures the effect of symbolic distance during all-pairs testing. This effect was consistently positive, and the 95% credible intervals exclude zero (Figure 4C and 4D) with performance increasing with larger symbolic distance. In the unequal reward condition, the SDE was robust for both within-list and between-lists pairs (Figure 4D, "All Within" vs. "All Between"). If anything, the between-lists SDE was slightly greater than that for within-list transfer. The SDE was stronger in the equal reward condition (Figure 4C) than it was in the unequal reward condition (Figure 4D), but was equally strong for all test pairs belonging to each reward condition. Overall, these results show a consistent and durable effect of symbolic distance, even when differential rewards produced strong biases in responding.

Figure 4E plots estimates of the reward disparity effect β_R in the unequal reward condition. β_R was close to 0.0 during training (Figure 4E "Adjacent Within Early" and "Adjacent Within Late"), suggesting that subjects did equally well on the training pairs for both lists even with reward disparity. However, a dramatic reward effect appeared when between-lists adjacent pairs were tested (Figure 4E, "Adjacent Between"), and this persisted throughout all-pairs testing (Figure 4E, "All Between"), indicating that subjects had a strong bias for the item drawn from the large reward list. Interestingly, even the within-list pairs have shown this effect during all pairs testing (Figure 4E "All Within"), reflecting higher performance on large reward pairs than small reward pairs, a result seemingly at odds with the comparable performance for both lists during the training phase. This last detail can't be explained by a mere "reward preference" and may signal a broader effect on motivation.

Figure 4F estimates the interaction of β_{DR} between distance β_D and reward β_R in the unequal reward condition. During within-list all-pairs testing, we observed a small positive interaction between distance and reward
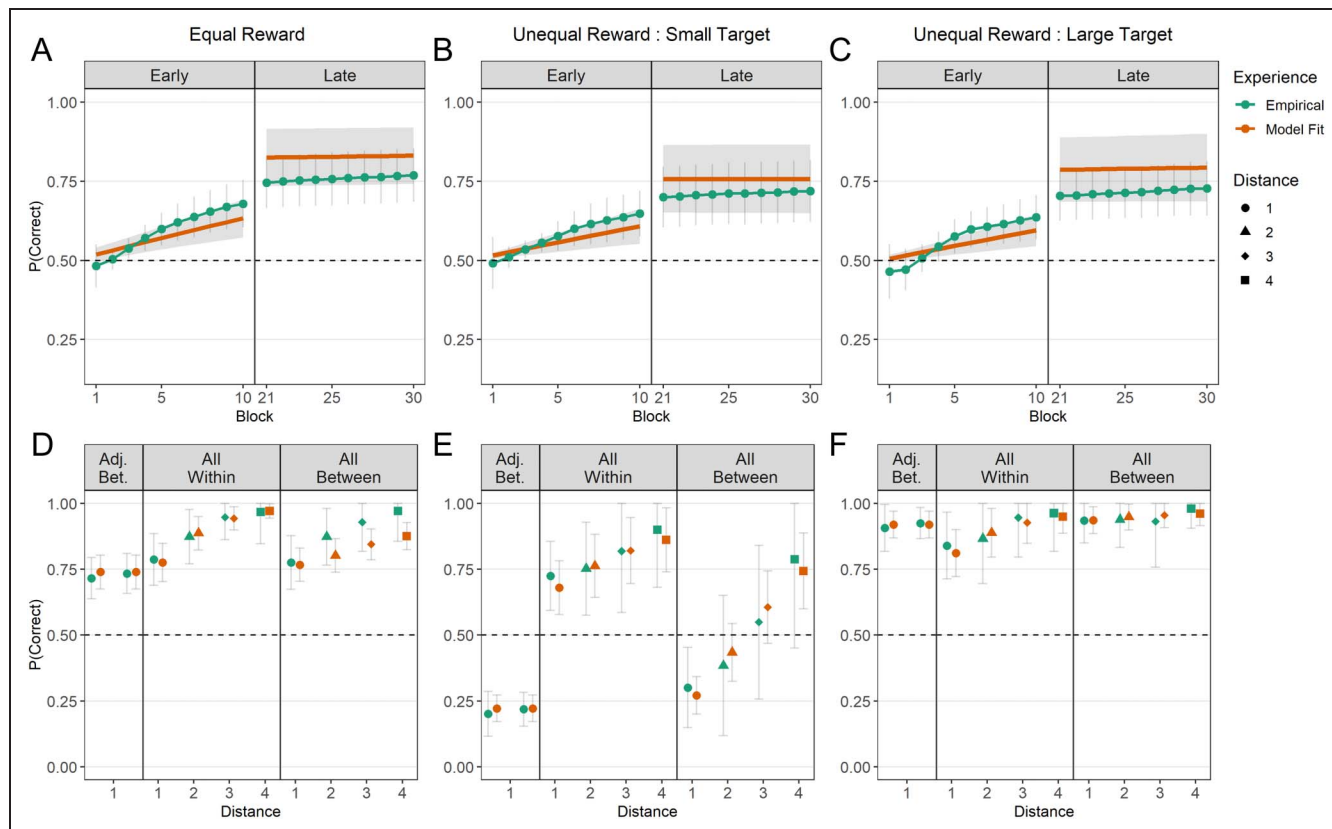

Figure 5. Empirical (green points) and model fit (orange lines and points) performance during training (A = equal, B = small reward, C = large reward) and testing (D = equal, E = small reward, F = large reward). Points represent average observed performance during each block, whereas line represents best-fit model estimates. Error bars and shaded regions depict 95% credible intervals. (A and D) Performance during equal reward presentations. (B and E) Performance when the targeted item was from the small reward list during unequal reward presentations. (C and F) Performance when the targeted item was from the large reward list during unequal reward presentations.

revealing that the distance effects were enhanced by reward magnitude (Figure 4F "All Within"). In contrast, we observed a small negative relationship during between-lists all-pairs testing revealing a competition between the distance effect and the reward effect, or a distance effect intensified by a small reward (Figure 4F "All Between"). However, both of these effects are also fairly uncertain, with posterior distributions that overlap with zero. As such, although the data are suggestive of this interaction, we cannot rule out a near-zero effect size for the interaction parameter.

Figure 5 compares the observed empirical means of pairs to the model fits calculated using the regression parameters described above. Figure 5A–5C shows the mean probability of a correct response during blocks of training trials for each condition during both early and late training. In general, observed data were in close correspondence with the model description, although some curvature evident in the first three or four blocks of the empirical data was not quite captured by the regression model.

A similarly good correspondence was found between empirical means and model fits of the testing phases. Above-chance responding and an evident SDE were captured in the equal reward condition (Figure 5D). Similarly, the interaction between symbolic distance and reward disparity yielded accurate recapitulation of both the interference effects when subjects were presented with small reward targets (Figure 5E) and the near-ceiling performance when subjects were presented with large-reward targets. In general, these results indicate that the parameters plotted in Figure 4 provide a good overall summary of the contributions of distance and reward effects, and as such give a suitable summary description of the data.

DISCUSSION

There is a long-standing controversy as to whether performance on serial learning tasks reflects the application of TI, or can instead be explained by model-free RL. We investigated the cognitive mechanisms underlying serial learning by putting reward-based and rule-based comparisons of stimuli into conflict with one another. The goal was to measure the relative influence of these two types of learning. It might have been the case that subjects always favored a rule-based strategy, even when it forced them to choose an item previously associated with a low reward. Conversely, subjects might have forsaken the application of transitivity and chosen only according to experienced reward associations.

Instead, the results revealed a complex interplay of rule- and reward-based strategies. When rewards were equal, it was clear that monkeys could apply the rule both within and across lists. However, when the stimulus was associated with a small reward, it was difficult for monkeys to choose that item when it was paired with a large-reward stimulus, even when the ordering rule that had been established during training dictated that they should do so. Such choices were suboptimal in that they resulted in no reward. However, subjects also displayed a critical hallmark of TI, the SDE, such that overcoming the biasing effect of differential rewards was easier when item ranks differed by a larger amount. RTs were much slower when making correct choices when these two influences were in conflict than when the influences were concordant, suggesting competition between rule- and reward-based decision strategies. The lack of a clear RT difference in the equal reward condition demonstrates that such large RT differences are not a typical feature of TI and thus point specifically to an impact of reward disparity on the decision-making process.

Before introducing between-lists test pairs with differential rewards, performance was broadly consistent with past reports of TI in rhesus macaques. When other factors were controlled for, subjects responded above chance overall at test and performed similarly on within- and between-lists pairs in most cases, suggesting that subjects did not find derived pairs difficult in general and that they relied on positional inference to solve novel pairings. It is noteworthy that the $\beta_\varnothing$ parameter was similar in both the equal and unequal reward conditions and reliably above chance following early training, suggesting that subjects displayed comparable overall learning whether or not a reward disparity was present during training (Figure 4A and 4C).

Subjects also displayed positive SDEs, as measured by the β_D parameter, in both equal and unequal reward conditions. Although the size of this effect was slightly smaller in the unequal reward condition, the effects were of a similar magnitude (Figure 4C and 4D). Thus, despite a powerful overall preference for stimuli taken from the large reward list (Figure 2B), subjects responded at greater than chance levels of accuracy to the most widely spaced pairs, as the SDE effectively overtook the reward bias for the largest symbolic distances. This result is consistent with past studies of TI, both those using standard paradigms and those using derived lists (Terrace, 2012; Jensen et al., 2020). The presence of both above-chance performance to test pairs and reliably positive SDEs point is not only evidence of TI, but, more broadly, it is evidence of a stable rule-based component of serial learning. This is in keeping with previous model-based proposals that serial learning relies on a representation of stimulus position along a continuum.

Interestingly, despite dramatic effects of reward disparity, the estimated intercept ($\beta_\varnothing$) and effect size of the SDE in our model (β_D) was similar in both the unequal and equal reward conditions, with the latter only slightly larger. However, despite estimates of $\beta_\varnothing$ and β_D providing clear evidence of TI, subjects also displayed a very large effect of reward disparity, as measured by the β_R parameter (Figure 4E). subjects displayed a strong preference for stimuli associated with the large reward list, resulting in enhanced accuracy presented with large reward targets,

but reduced (and often below-chance) accuracy when presented with large reward distractors (Figure 2B and 2C, "Adj. Bet." and "All Between"). Although these effects "average out" to above-chance performance overall, the effect of differential reward still dominates preference in most cases. Furthermore, discrepancies in RT between correct and incorrect responses in the unequal reward condition (Figure 3B–3C) suggest a competition between rule-based and reward-based strategies, validated by the analysis of interaction between distance and reward (Figure 4F), that were cognitively costly when the "target" (according to the rule) was associated with small rewards.

The effect size of the reward manipulation was much larger than that reported by Jensen, Alkan, et al. (2019), who also put reward discrepancies in conflict with the ordinal rank of list items. subjects may have had more success overcoming the reward bias in that study because, unlike this study, they were trained on single lists with a unique reward for each stimulus. Thus, the derived list paradigm (and thus using positional inference to combine knowledge across multiple lists) may be more vulnerable to this kind of disruption than TI alone. Another possible factor is that subjects in this study received less training for each list than was the case in the 2019 experiment, where each list was trained for multiple sessions before transfer.

Furthermore, the effect of the reward was also a positive within-list all-pairs testing (Figure 4E), resulting in lower within-list performance for test items from the small reward list. Because within-list small reward stimuli did not have reward discrepancies with respect to one another, this must reflect some broader effect on learning induced by reward disparities across trials. For example, subjects may have exhibited differential motivation for (or paid differential attention to) stimuli from one list relative to the other. Reward magnitude has played an important role in modulating the learning rate during other tasks (Ferrucci et al., 2019; Vartak, Jeurissen, Self, & Roelfsema, 2017) and thus may do more than merely increase a response bias. Thus, learning was modulated by the amount of reward, impacting both motivation and attention, and doing so in a more sophisticated manner than mere reward seeking (reviewed by Jovanovic & Matejevic, 2014).

Following these interpretations, the β_R observed during between-lists presentations could be a mixture of (a) variations in attention and motivation that influenced learning and (b) a response bias arising directly from the noninformative reward disparity itself. At a minimum, a purely reward-associative account would only predict an increased β_R during between-lists presentations. As such, this differential learning may be another clue as to why subsequent between-lists reward biases were as large as were observed. A future study with a more direct index of either motivation or attention could clarify to what extent these are modulated by differential rewards.

Although differential rewards had a powerful influence on preference, a strictly model-free account is not sufficient to explain the behavior observed in this study. Both the consistently positive SDEs and greater-than-zero intercept terms suggest that model-based learning was active and stable during training and, to some extent, independent of reward associations. Rhesus macaques appear, on average, to favor rule-based choices even when the corresponding reward sizes conflict with the rule (see also the work of Gazes et al., 2012). We interpret our results as suggesting that the representational strategy was applied by subjects in both conditions and, to similar degrees, independent of the confounding effect of differential rewards. However, based on our analysis of RT and predictors, the influence of reward on accuracy seems to be somewhat in competition with the rule-based representation, rather than reward value being fully and directly integrated into each subject's representation of order. Rule-based and reward-based contributions having a degree of independence may also explain why such consistent distance effects (as measured by β_D) were observed, even as performance on specific pairs was disrupted.

Our results make clear that a full picture of learning must incorporate reward-based associative effects, while retaining the transitive characteristic of a cognitive model. Although two competing learning systems may be at play, a theory of TI should ultimately incorporate both effects, rather than treating these factors as each being evidence for mutually exclusive theories of behavior (Collins & Cockburn, 2020).

Our present results are also consistent with previous demonstrations that subjects rely on positional inference when evaluating between-lists pairs (Jensen et al., 2020; Terrace, 2012), doing so to a similar degree both in the presence and absence of differential reward effects. Thus, we attribute the low response accuracy for between-lists critical pairs during the small reward trials (Figure 2D–2G) as being because of reward bias instead of errors in positional inference. The presence of a positive SDE in between-lists all-pairs presentations (Figure 2B, 4C, and 4D) supports the understanding of the rules and the presence of a reward bias.

Overall, we observed that monkeys seem to use both model-free and model-based RL to perform transitive and positional inference, in a specific context of dual-list TI with reward disparity. The reward effect seems to interact with the representation of list order and transitive rule according to the symbolic distance, generating an improvement of learning and better performance for the large reward list than is observed during classic TI performance (i.e., the equal reward condition). However, when ordinal position and corresponding reward each favor a conflicting alternative, as in between-lists pairs when the targeted item is from the small reward list, the two types of learning, model-free and model-based, are also in conflict relative to both the amount of reward and the informativeness of the rule (in this case, the symbolic distance between items). This conflict is evidenced by the presence of both clear reward biases and of differential RTs as a

function of conflict between these factors. These findings point to considerations that should be explored using TI tasks that employ derived lists. Because performance reflects the influence of both rules and reward, an important objective is to better understand the ways in which these strategies interact in serial learning.

Acknowledgments

Reprint requests should be sent to Vincent P. Ferrera, Zuckerman Mind Brain Behavior Institute, Columbia University, 3227 Broadway, New York, NY, 10027, or via e-mail: vpf3@cumc .columbia.edu.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .586, W/M = .207, M/W = .172, and W/W = .036.

REFERENCES

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436. https://doi.org/10.1163/156856897X00357

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Chen, S., Swartz, K. B., & Terrace, H. S. (1997). Knowledge of the ordinal position of list items in rhesus monkeys. *Psychological Science*, *8*, 80–86. https://doi.org/10.1111/j .1467-9280.1997.tb00687.x

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*, 576–586. https://doi.org/10.1038/s41583-020-0355-6, PubMed: 32873936

D'Amato, M. R., & Colombo, M. (1990). The symbolic distance effect in monkeys (*Cebus apella*). *Animal Learning & Behavior*, *18*, 133–140. https://doi.org/10.3758/BF03205250

Davis, H. (1992). Transitive inference in rats (*Rattus norvegicus*). *Journal of Comparative Psychology*, *106*, 342–349. https://doi .org/10.1037/0735-7036.106.3.342, PubMed: 1451416

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 1–12. https://doi.org/10.1016/j.conb.2008.08.003, PubMed: 18708140

Ferrucci, L., Nougaret, S., Brunamonti, E., & Genovesio, A. (2019). Effects of reward size and context on learning in macaque monkeys. *Behavioural Brain Research*, *372*, 111983. https://doi.org/10.1016/j.bbr.2019.111983, PubMed: 31141723

Gazes, R. P., Chee, N. W., & Hampton, R. R. (2012). Cognitive mechanisms for transitive inference performance in rhesus monkeys: Measuring the influence of associative strength and inferred order. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 331–345. https://doi.org/10.1037 /a0030306, PubMed: 23066978

Gillan, D. J. (1981). Reasoning in the chimpanzee: II. Transitive inference. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*, 150–164. https://doi.org/10.1037/0097 -7403.7.2.150

Hakes, D. T., James, C. T., & Young, R. K. (1964). A re-examination of the Ebbinghaus derived-list paradigm. *Journal of Experimental Psychology*, *68*, 508–514. https://doi .org/10.1037/h0046779, PubMed: 14212138

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*, 497–505. https://doi.org/10.1016/j.tics .2010.08.005, PubMed: 20884275

Jensen, G. (2017). Serial learning. In J. Call, G. M. Bughardt, I. M. Pepperberg, C. T. Snowdon, T. Zentall (Eds.), *APA handbook of comparative psychology: Perception, learning, and cognition* (pp. 385–409). American Psychological Association. https://doi.org/10.1037/0000012-018

Jensen, G., Alkan, Y., Ferrera, V. P., & Terrace, H. S. (2019). Reward associations do not explain transitive inference performance in monkeys. *Science Advances*, *5*, eaaw2089. https://doi.org/10.1126/sciadv.aaw2089, PubMed: 32128384

Jensen, G., Ferrera, V. P., & Terrace, H. S. (2020). Positional inference in rhesus macaques. *PsyArXiv Preprints*. https://doi .org/10.31234/osf.io/pe5rm

Jensen, G., Terrace, H. S., & Ferrera, V. P. (2019). Discovering implied serial order through model-free and model-based learning. *Frontiers in Neuroscience*, *13*, 878. https://doi.org /10.3389/fnins.2019.00878, PubMed: 31481871

Jovanovic, D., & Matejevic, M. (2014). Relationship between rewards and intrinsic motivation for learning—Researches review. *Procedia - Social and Behavioral Sciences*, *149*, 456–460. https://doi.org/10.1016/j.sbspro.2014.08.287

Kao, T., Jensen, G., Michaelcheck, C., Ferrera, V. P., & Terrace, H. S. (2020). Absolute and relative knowledge of ordinal position on implied lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*, 2227–2243. https://doi .org/10.1037/xlm0000783, PubMed: 31750719

Kleiner, M., Brainard, D. H., Pelli, D. G., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, 1–16.

Lazareva, O. F., Smirnova, A. A., Zorina, Z. A., & Rayevsky, V. V. (2001). Hooded crows solve a transitive inference problem cognitively. *Animal Welfare*, *10*, 219–231.

Lazareva, O. F., & Wasserman, E. A. (2006). Effect of stimulus order ability and reinforcement history on transitive responding in pigeons. *Behavioural Processes*, *72*, 161–172. https://doi.org/10.1016/j.beproc.2006.01.008, PubMed: 16460886

Lazareva, O. F., & Wasserman, E. A. (2012). Transitive inference in pigeons: Measuring the associative values of stimuli B and D. *Behavioural Processes*, *89*, 244–255. https://doi.org/10 .1016/j.beproc.2011.12.001, PubMed: 22178714

McGonigle, B. O., & Chalmers, M. (1977). Are monkeys logical? *Nature*, *267*, 694–696. https://doi.org/10.1038/267694a0, PubMed: 406574

Merritt, D. J., & Terrace, H. S. (2011). Mechanisms of inferential order judgments in humans (*Homo sapiens*) and rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology*, *25*, 227–238. https://doi.org/10.1037/a0021572, PubMed: 21341909

Minamimoto, T., La Camera, G., & Richmond, B. J. (2009). Measuring and modeling the interaction among reward size, delay to reward, and satiation level on motivation in monkeys. *Journal of Neurophysiology*, *101*, 437–447. https://doi.org/10.1152/jn.90959.2008, PubMed: 18987119

Mione, V., Brunamonti, E., Pani, P., Genovesio, A., & Ferraina, S. (2020). Dorsal premotor cortex neurons signal the level of choice difficulty during logical decisions. *Cell Reports*, *32*, 107961. https://doi.org/10.1016/j.celrep.2020.107961, PubMed: 32726625

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. https://doi.org/10.1163/156856897X00366, PubMed: 9176953

Piaget, J. (1921). Une forme verbale de la comparison chez l'enfant. *Archives de Psychologie*, *18*, 141–172.

Smith, J. D., & Church, B. A. (2018). Dissociable learning processes in comparative psychology. *Psychonomic Bulletin & Review*, *25*, 1565–1584. https://doi.org/10.3758/s13423-017-1353-1, PubMed: 28799092

Stan Development Team. (2021). *Stan modeling language users guide and reference manual*, 2.28. https://mc-stan.org

Staněk, R. (2017). Home bias in sport betting: Evidence from the Czech betting market. *Judgment and Decision-making*, *12*, 168–172.

Stanisor, L., Van Der Togt, C., Pennartz, C. M. A., & Roelfsema, P. R. (2013). A unified selection signal for attention and reward in primary visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *110*, 9136–9141. https://doi.org/10.1073/pnas.1300117110, PubMed: 23676276

Swartz, K. B., Chen, S., & Terrace, H. S. (1991). Serial learning by rhesus monkeys: I. acquisition and retention of multiple four-item lists. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*, 396–410. https://doi.org/10.1037/0097-7403.17.4.396, PubMed: 1744595

Terrace, H. S. (2010). The comparative psychology of serially organized behavior. *Comparative Cognition & Behavior Reviews*, *5*, 23–58. https://doi.org/10.3819/ccbr.2010.50002

Terrace, H. S. (2012). The comparative psychology of ordinal knowledge. In T. R. Zentall & E. A. Wasserman (Eds.), *The Oxford handbook of comparative cognition* (pp. 615–651). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195392661.013.0032

Treichler, F. R., & Raghanti, M. A. (2010). Serial list combination by monkeys (*Macaca mulatta*): Test cues and linking. *Animal Cognition*, *13*, 121–131. https://doi.org/10.1007/s10071-009-0251-y, PubMed: 19533183

Treichler, F. R., & Van Tilburg, D. (1996). Concurrent conditional discrimination tests of transitive inference by macaque monkeys: List linking. *Journal of Experimental Psychology: Animal Behavior Processes*, *22*, 105–117. https://doi.org/10.1037/0097-7403.22.1.105, PubMed: 8568492

Vartak, D., Jeurissen, D., Self, M. W., & Roelfsema, P. R. (2017). The influence of attention and reward on the learning of stimulus-response associations. *Scientific Reports*, *7*, 9036. https://doi.org/10.1038/s41598-017-08200-w, PubMed: 28831043

Vasconcelos, M., & Monteiro, T. (2014). European starlings unriddle the ambiguous-cue problem. *Frontiers in Psychology*, *5*, 944. https://doi.org/10.3389/fpsyg.2014.00944, PubMed: 25206346

von Fersen, L., Wynne, C. D., Delius, J. D., & Staddon, J. E. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*, 334–341. https://doi.org/10.1037/0097-7403.17.3.334